UNetX: Real-time Pedestrian Crosswalk Segmentation on Mobile Device

Eduard Adaska
Computer Systems Department
Belarusian State University
of Informatics and Radioelectronics
Minsk, Belarus

Anton Lechanka
Computer Systems Department
Belarusian State University
of Informatics and Radioelectronics
Minsk, Belarus
lechenko@ bsuir.by

Abstract. This paper presents a lightweight deep neural network that segments pedestrian crosswalks on an image in realtime. It is based on U-Net architecture with all its convolution layers substituted with depthwise separable convolution ones. This neural network was trained and tested against a set of manually segmented 3083 road images — with and without crosswalks. The resulting network has only 383K parameters and runs at 35 FPS on a mobile phone. The Jaccard index (IoU metric) on the validation set is 0.9138.

Keywords: **autonomous car, pedestrian crosswalk, image segmentation, deep neural network, U-Net, depthwise separable convolution**

I. INTRODUCTION

There have been considerable research interests in autonomous cars in the past few years, primarily concerning the safety of autonomous vehicles. Meanwhile, Every year, approximately 300,000 pedestrians die on the roads accounting for up to 26% of all deaths in road accidents [1]. That makes pedestrian crosswalk detection an essential task for the safety of autonomous cars.

The main three approaches for crosswalk detection are based only on camera images [2, 3], only on LIDAR point clouds [4] and on both images and point clouds [5, 6]. In this text, we present a convolutional neural network, UNetX, based entirely on camera image processing. The neural network can be used as a part of a complex autonomous vehicle system. Nevertheless, being a lightweight solution that can run in realtime on a mobile phone, it also can operate as a virtual offline road assistant for drivers.

This network solves the detection problem as image segmentation. It partitions an image into two segments: the one that belongs to a crosswalk and the one that does not. In this paper, we present UNetX, a five-time more compact modification of U-Net [7], which makes UNetX faster and more power-efficient.

II. METHODOLOGY

A. Data set selection and annotation

The data set of 35625 images for crosswalk detection [8] have been taken for that problem. That dataset effectively comprises three data subsets. The first data subset consists of images captured using a Bumblebee XB3 camera mounted on the IARA vehicle during the day. The second one was created using the same instrument at night. And the third subset was taken with GoPro HERO 3 camera in Full HD.

The set was annotated only with crosswalk presence property; thus, additional annotation was required. A portion of those images has been manually annotated using VGG Image Annotator [9]. JSON output from VGG Image Annotator was converted to a bit-mask representing whether the pixel belongs to a crosswalk. Several local streets images have also been added. Those images have been taken during both day and night and also during bad weather. The images have been evenly selected from all data subsets. In total, 3083 images have been annotated, 975 of which did not contain any crosswalk.

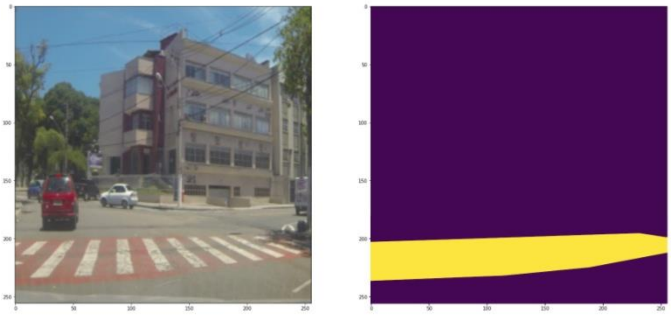

Fig. 1. Example of an image on the left and its annotation mask on the right

A benchmark application was developed to measure realtime performance on a mobile device. It is an iOS application that runs the CNN model using

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

CoreML framework [10]. It feeds a batch of images to a model, measures their processing time, and calculates frames per second (FPS). An iPhone 8 was used to measure real-time performance on an actual device.

B. *CNN architecture*

The U-Net was used as a base convolutional network with slight changes. Unlike the original paper, in this work, 256 × 256 image with 3 color channels is used as an input for UNet CNN. An encoder transforms an input image to a tensor 64×16×16, and a decoder with concatenation layers unpack this tensor to a single channel 256 × 256 mask. Each value of a mask is a probability that the corresponding pixel belongs to a pedestrian crosswalk on an input image.

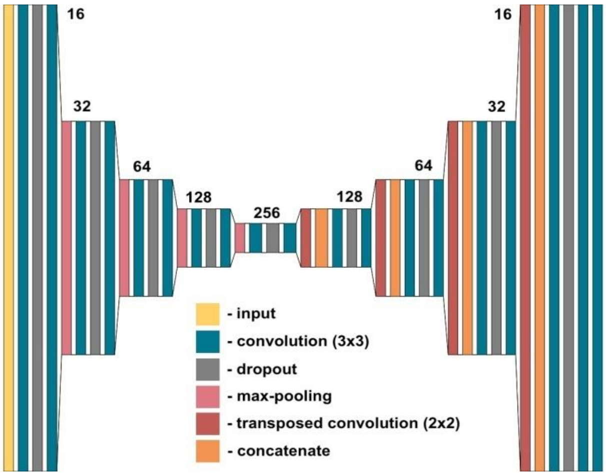

Fig. 2. U-Net CNN visualization

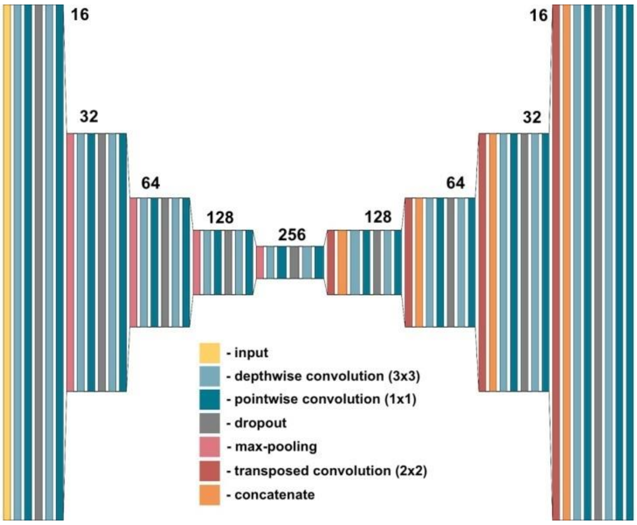
16
32
64
128
256
128
64
32
16
🟨 - input
🟦 - depthwise convolution (3x3)
🟦 - pointwise convolution (1x1)
⬛ - dropout
🟥 - max-pooling
🟥 - transposed convolution (2x2)
🟧 - concatenate

Fig. 3. UNetX CNN visualization

The U-Net network has only 1,941,105 parameters, but even a smaller network has been proposed. Convolution layers in U-Net have been substituted with depthwise separable convolution layers [11]. The other components of a CNN, like max-pooling and concatenation layers, are unchanged.

Depthwise separable convolution is essentially two convolutions. The first one is a depthwise convolution. It takes N×M×C tensor and performs C 2D convolutions with its filter on each N×M layer. The outputs of a depthwise convolution are stacked to a 3D tensor and fed to a point-wise convolution. This modification results in a five times smaller network with only 383K parameters.

C. *Figure of merit*

Intersection over Union (IoU) is used as an evaluation metric for the CNNs described. It is a commonly used performance metric for image segmentation problems [12].It measures the similarity between a predicted region and a ground truth region. The predicted region is defined by the output of a CNN with a probability higher than 0.5, and the ground truth is an annotation mask. In order to calculate IoU, one needs to find an intersection and a union of the predicted region and the ground truth. Thus, the IoU is a ratio of the area of intersection over the area of union.

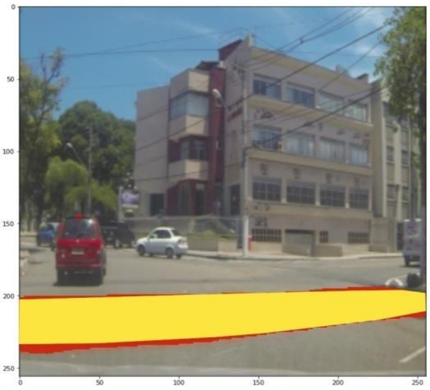

Fig. 4. Intersection (yellow) over Union (red and yellow) visualization

III. TRAINING THE MODEL

Both UNet and UNet-X CNNs have been trained on NVIDIA Tesla V100 using Keras [13]. Each network was trained for 100 epochs; each epoch processed the whole training set divided into batches of 32. The optimization algorithm is Adam [14]. The learning rate and other training hyper-parameters were set to Keras defaults.

Binary cross-entropy [15] was used as a loss function during training, as in:

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(p(y_i)) + (1-y_i)\log(1-p(y_i)) \quad (1)$$

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

N in (1) is the number of pixels, yiis the ground truth value whether a pixel number ibelongs to a pedestrian crosswalk segment, and p(yi) is a predicted probability that a given pixel belongs to the segment.

The data set has been randomly split into training and validation sets. The training set comprises 2159 images (70% of the annotated data set), and the validation set is 924 images. After each single training epoch is concluded, the trained network model is validated against the validation set. Fig. 5 shows model training loss and validation IoU after each epoch.

IV. RESULTS AND DISCUSSION

Substitution of convolution layers with depthwise convolutions insignificantly affects segmentation results. Values of the IoU metric given in a table I and the visual comparison of the segmentation masks, as shown in fig. 6, support this claim. At the same time, the number of trained network parameters decreased five times, from 1,941,105 to 386.543. This improved network performance on a mobile device by 33% up to 35 frames per second, which is enough to perform proper real-time segmentation.

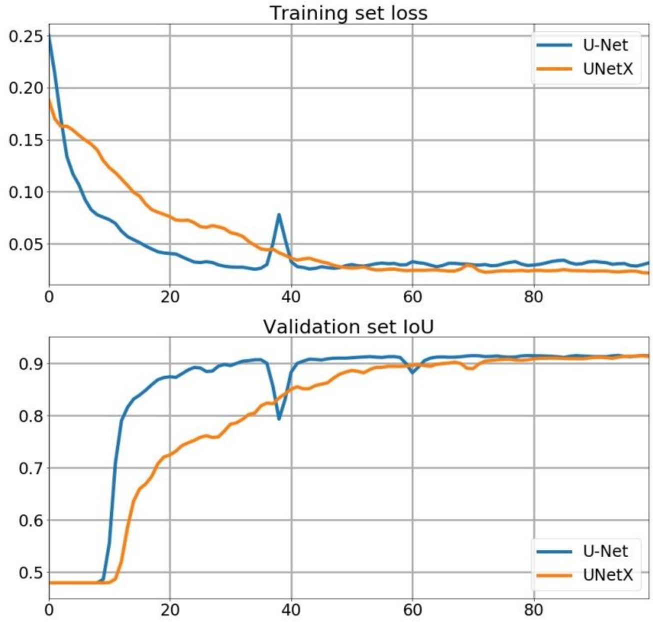

Fig. 5. Loss and IoU training progress

Despite the dramatic decrease in computation workload, the performance improvement is modest. This is due to the almost unchanged number of load/store operations per inference and the increased number of CNN layers. Performance results are summarized in Table I.

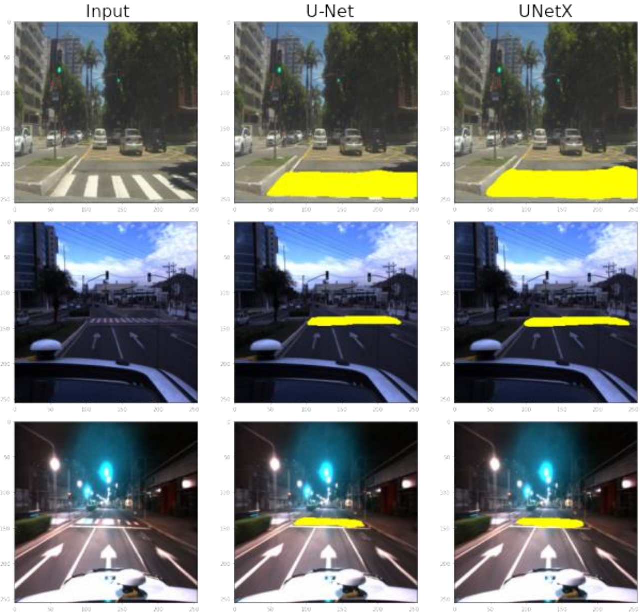

Fig. 6. Visual comparison

TABLE I. MODELS SUMMARY

Model	*# of parameters*	*IoU*	*FPS*
U-Net	1,941,105	0.9147	26.31
UNetX	386,543	0.9138	35.17

V. CONCLUSION

Encoder-Decoder CNN with depthwise separable convolutions can be used as a tool for real-time segmentation of pedestrian crosswalks. Moreover, according to the conducted experiment, switching to depthwise separable convolution has an insensible impact on segmentation quality. In conclusion, the proposed UNetX can be used as a part of a complex system of an autonomous car or as a virtual assistant for drivers, thanks to excellent real-time performance on mobile devices.

REFERENCES

[1] W. H. Organization et al., “Global status report on road safety 2018: summary,” World Health Organization, Tech. Rep., 2018.

[2] M. A. Malbog, “Mask r-cnn for pedestrian crosswalk detection and instance segmentation,” in 2019 IEEE 6th International Conference on Engineering Technologies and Applied Sciences (ICETAS). IEEE, 2019, pp. 1–5.

[3] R. Benenson, M. Mathias, R. Timofte, and L. Van Gool, “Pedestrian detection at 100 frames per second,” in 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012, pp. 2903–2910.

[4] H. Wang, B. Wang, B. Liu, X. Meng, and G. Yang, “Pedestrian recognition and tracking using 3d lidar for autonomous vehicle,” Robotics and Autonomous Systems, vol. 88, pp. 71–78, 2017.

[5] R. Guidolini, L. G. Scart, L. F. R. Jesus, V. B. Cardoso, C. Badue, and T. Oliveira-Santos, “Handling pedestrians in crosswalks using deep neural networks in the iara

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

autonomous car," in 2018 International Joint Conference on Neural Networks (IJCNN), 2018, pp. 1–8.

[6] J. Dou, J. Fang, T. Li, and J. Xue, "Boosting cnn-based pedestrian detection via 3d lidar fusion in autonomous driving," in International Conference on Image and Graphics. Springer, 2017, pp. 3–13.

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, pp. 234–241.

[8] R. F. Berriel, F. S. Rossi, A. F. de Souza, and T. Oliveira-Santos, "Automatic large-scale data acquisition via crowdsourcing for crosswalk classification: A deep learning approach," Computers & Graphics, vol. 68, pp. 32–42, 2017.

[9] A. Dutta, A. Gupta, and A. Zissermann, "Vgg image annotator (via)," http://www. robots. ox. ac. uk/vgg/ software/via, 2016.

[10] M. Thakkar, Beginning machine learning in iOS: CoreML Framework. Apress, 2019.

[11] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1251–1258.

[12] M. A. Rahman and Y. Wang, "Optimizing intersection-over-union in deep neural networks for image segmentation," in International symposium on visual computing. Springer, 2016, pp. 234–244.

[13] A. Gulli and S. Pal, Deep learning with Keras. Packt Publishing Ltd, 2017.

[14] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[15] S. Jadon, "A survey of loss functions for semantic segmentation," in 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). IEEE, 2020, pp. 1–7.

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.